MINISTRY OF EDUCATION

BELARUSIAN STATE UNIVERSITY

NATIONAL RESEARCH CENTER FOR APPLIED PROBLEMS
OF MATHEMATICS AND INFORMATICS

SCHOOL OF APPLIED MATHEMATICS AND INFORMATICS

BELARUSIAN REPUBLICAN FOUNDATION
FOR FUNDAMENTAL RESEARCH

BELARUSIAN STATISTICAL ASSOCIATION

COMPUTER DATA ANALYSIS AND MODELING

COMPLEX STOCHASTIC DATA AND SYSTEMS

Proceedings of the Eighth International Conference
Minsk, September 11-15
Volume 1

MINSK
"Publishing center BSU"
2007

UDC 519.24(063)+519.95(063)

EDITORS:
Prof. Dr. S. Aivazian, Prof. Dr. P. Filzmoser, Prof. Dr. Yu. Kharin

Computer Data Analysis and Modeling: Complex Stochastic Data and Systems: Proc. of the Eighth International Conference (September 11-15, 2007, Minsk). Vol. 1. Minsk: Belarusian State University, 2007. — 283 p.
ISBN

This collection of papers includes proceedings of the Eighth International Conference "Computer Data Analysis and Modeling: Complex Stochastic Data and Systems" organized by the Belarusian State University and held in September 2007 in Minsk. Papers are reviewed by qualified researchers from Belarus, Russia, Austria, Great Britain, Germany, Lithuania.

The papers are devoted to the topical problems: robust and nonparametric data analysis; statistical analysis of time series and forecasting; multivariate data analysis; design of experiments; probability and statistical analysis of discrete data; econometric analysis and modeling; survey analysis and official statistics; computer intensive methods, algorithms and software; computer data analysis in applications.

For specialists who work in the fields of mathematical statistics and its applications, computer data analysis, statistical modeling and statistical software development.

ISBN

PREFACE

The Eighth International Conference "Computer Data Analysis and Modeling: Complex Stochastic Data and Systems" (CDAM'2007) organized by the Belarusian State University on September 11-15, 2007, was devoted to the topical problems in computer data analysis and modeling. Statistical methods of computer data analysis and modeling are widely used in variety of fields: computer support of scientific research; decision making in economics, business, engineering, medicine end ecology; statistical modeling of complex systems of different nature and purpose. In the Republic of Belarus computer data analysis and modeling have been developed successfully for the last 25 years. Scientific conferences CDAM were held in September 1988, December 1990, December 1992, September 1995, June 1998, September 2001 and September 2004 in Minsk.

The Proceedings of the CDAM'2007 include 2 volumes containing 107 papers. The topics of the papers correspond to the following scientific problems: robust and nonparametric statistical analysis of time series and forecasting; multivariate data analysis; statistical classification and pattern recognition; signal processing; statistical modeling; modeling of complex systems in different applications; statistics in economics, finance and other fields, software for data analysis and statistical modeling.

The Organizing Committee of the CDAM'2007 makes its acknowledgements to: the Belarusian State University; the National Research Center for Applied Problems of Mathematics and Informatics; the Belarusian Republican Foundation for Fundamental Research; Belarusian Statistical Association; School of Applied Mathematics and Informatics; the Vienna University of Technology; BelSwissBank for financial support.

S. Aivazian
P. Filzmoser
Yu. Kharin

CONTENTS

PLENARY LECTURES

PLENARY LECTURES

APPROXIMATE ENVELOPES FOR FINDING AN UNKNOWN NUMBER OF MULTIVARIATE OUTLIERS IN LARGE DATA SETS

A.C. Atkinson*, M. Riani† and F. Laurini‡
*London School of Economics
London WC2A 2AE, U.K.
e-mail: a.c.atkinson@lse.ac.uk
†‡ Università di Parma
Parma 43100, Italy
†e-mail: mriani@unipr.it
‡e-mail: fabrizio.laurini@unipr.it

Abstract

We provide thresholds for the test statistic for multiple outliers in multivariate normal samples. Except in small problems, direct simulation of the required combinations of sample size, number of outliers, dimension of the observations and significance level is excessively time consuming. We find the thresholds by rescaling a paradigmatic curve found by simulation. Our method is illustrated on an example with 1,827 observations.

1 Introduction

The normal distribution has a central place in the analysis of multivariate data. Unfortunately the moment estimates of the mean and covariance matrix are extremely sensitive to the presence of outliers.

The standard procedure for detecting outliers is to select m observations, known to be outlier free, from the total of n observations, and to obtain from them "good", that is uncontaminated, estimates of the parameters. In this paper we use the forward search, [3], to provide robust ("good") estimates of the parameters. Outliers are detected by tests of the $(m+1)$st ordered Mahalanobis distances based on these robust estimates.

For data with "small" n, perhaps up to 200, and of moderate dimension v, we can use direct simulation to find the null distribution of the outlier test. But as v and n increase we need approximate methods; in the supermarket data introduced in §3.1 $n = 1{,}827$ and $v = 10$. In this paper we provide easily calculated approximations to the simulated percentage points of this outlier test. As our example shows, we typically need to apply the bounds for different values of n as outliers are detected. Straightforward interpolation in the simulation results is therefore essential.

Results on the distribution of deletion Mahalanobis distances in the absence of contamination include those of [10], [5] and Chapter 2 of [3]. There is a large literature on robust estimation and outlier detection including [9], [4], [11], [7] and [8]. Forward procedures were introduced by [6] and [1].

The Mahalanobis distances and outlier test are described more formally in §2. The supermarket data are introduced in §3.1. Section 3.2 illustrates the structure of the

bounds for different sample sizes and dimensions of data. The succeeding section provides simply calculated approximations to the threshold of the test. Our analysis of the supermarket data is in §5 with the data in Appendix 1. A technical result on Mahalanobis distances is in Appendix 2.

2 Distances

The squared Mahalanobis distances for the sample are defined as

$$d_i^2 = \{y_i - \hat{\mu}\}^T \hat{\Sigma}^{-1} \{y_i - \hat{\mu}\}, \qquad i = 1, \ldots, n, \tag{1}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the unbiased moment estimators of the mean and covariance matrix of the n observations and y_i is $v \times 1$.

In the forward search the parameters μ and Σ are estimated in the same way from a subset of m observations, yielding estimates $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$. From this subset we obtain n squared Mahalanobis distances

$$d_i^2(m) = \{y_i - \hat{\mu}(m)\}^T \hat{\Sigma}^{-1}(m) \{y_i - \hat{\mu}(m)\}, \qquad i = 1, \ldots, n. \tag{2}$$

We start with a subset of m_0 observations which grows in size during the search. When a subset $S(m)$ of m observations is used in fitting, we order the squared distances and take the observations corresponding to the $m + 1$ smallest as the new subset $S(m + 1)$.

In our examples we look at forward plots of quantities derived from the distances $d_i(m)$. These distances for $i \notin S(m)$ tend to decrease as n increases. If interest is in the latter part of the search we may use **scaled** distances

$$d_i^{\mathrm{sc}}(m) = d_i(m) \times \left(|\hat{\Sigma}(m)| / |\hat{\Sigma}(n)| \right)^{1/2v}, \tag{3}$$

where $\hat{\Sigma}(n)$ is the estimate of Σ at the end of the search.

To detect outliers we examine the minimum Mahalanobis distance amongst observations not in the subset

$$d_{\min}(m) = \min d_i(m) \quad i \notin S(m), \tag{4}$$

or its scaled version $d_{\min}^{\mathrm{sc}}(m)$. If this observation is an outlier relative to the other m observations, this distance will be large compared to the maximum Mahalanobis distance of observations in the subset.

3 The Structure of Forward Plots

3.1 Supermarket Data

Appendix 1 describes data from an Italian supermarket chain with n = 1,827 and v = 10. The 10 variables are measures of expenditure and purchasing behaviour, which

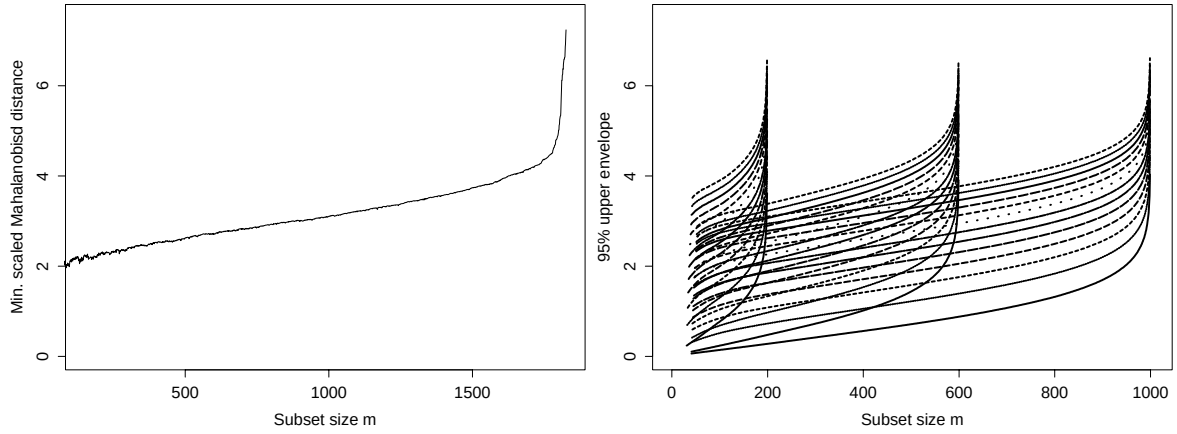

Figure 1: Forward plots of scaled minimum Mahalanobis $d_{\min}(m)$. Left-hand panel, supermarket data, right-hand panel 95% points from simulations as v goes from 1 to 13. Small v at the bottom of the plot: $n = 200, 600$ and 1,000

have been transformed by taking logs. These data are sufficiently large for straight simulation to be inconveniently lengthy.

Forward plots of the scaled minimum Mahalanobis distances are in the left-hand plot of Figure 1. It is clear from the large values, around 7, of the distances at the end of the search that there are some outliers. How many is not clear as, even with a normal sample, we would expect a few distances at the end of the search to be somewhat larger than the rest. Are these observations more remote than would be expected from the order statistics for the distribution of the minimum Mahalanobis distance?

3.2 Plots

The right-hand panel of Figure 1 shows forward plots of simulation envelopes for scaled minimum Mahalanobis distances from 10,000 simulations for sample sizes $n = 200$, 600 and 1,000. The envelope given is the 95% point of the empirical distribution of the minimum Mahalanobis distance amongst observations not in the subset for v from 1 to 13. There is clearly much common structure as n and v vary. Although the curves are all for sample sizes appreciably smaller than the 1,827 observations of the supermarket data, all have a maximum only around six. In general the plots look like a series of superimposed prows of viking longships. Local polynomial approximations to these curves were used by [2] for moderate sized samples. Here we exploit this common structure by rescaling a paradigmatic curve from a single simulation to find approximate envelopes for large n and v.

The envelopes for scaled distances are less curved than those for unscaled distances and so are easier to approximate. For brevity, only the results for scaled distances are presented here.

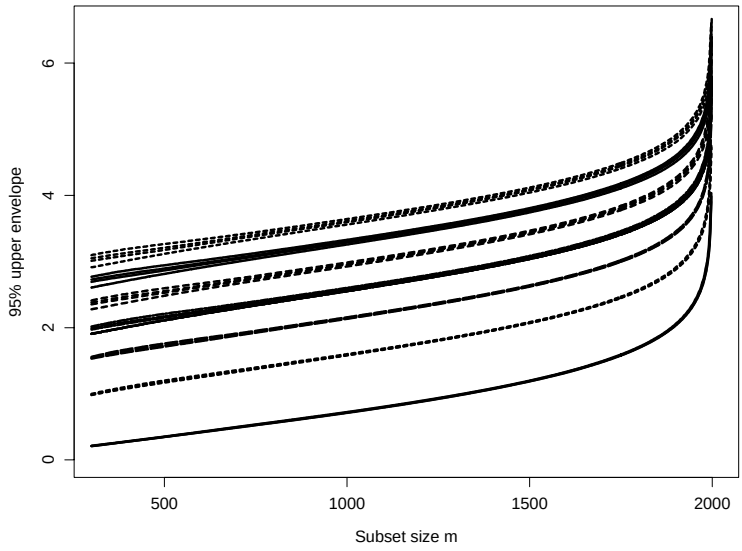

Figure 2: Scaled minimum Mahalanobis distances: 95% point for $v = 1$, 3, 5, 7, 9, 11 and 13 rescaled to have a common ordinate $x = 1999m/(n-1)$; $n = 600$, 800, 1,000 and 2,000. In each group of four curves, the smallest value of n is at the top

4 Rescaling a Paradigmatic Curve

For large data sets we use a simulation with $n = 2{,}000$ to provide our paradigmatic curve. The plots of distances in the right-hand panel of Figure 1 suggest that the common structure of the curves varies with v but hardly with m/n. To find an analytical form for this structure we first rescale all curves as a function of $x = m/(n-1)$; here $x = 1999m/(n-1)$. The resulting plot for scaled distances is in Figure 2 for six values of v from 1 to 13 and for $n = 600$, 800, 1,000 and 2,000. The major differences between the curves are for values of v, with a slight effect of n.

To find the effect of n we looked at the differences between the curves for two different values of n for various v. Since the curves have differing ordinates, we used linear interpolation to provide values at the points of the curve for $n = 2{,}000$. The left-hand panel of Figure 3 shows the resulting differences between the curves for $n = 2{,}000$ and $n = 1{,}000$, with large values of v at the bottom of the plot. The differences are virtually constant from $m = 500$ until close to the end of the search.

To use this structure we calculate an average value of the difference between $n = 2{,}000$ and the curves for several other values of n for $v = 13$. Plotting these differences against n gives a linear extrapolation in n for scaled and unscaled distances. We use this linear extrapolation for all v. Our comparisons show that we obtain good approximations over the whole range of m, even down to $n = 400$.

We also need to approximate the more important dependence on v. From Figure 2 we take the average, over the four values of n, of the curves for each v. We then subtract the curve for $v-1$ from that for v to obtain the 12 curves of differences shown in the right-hand panel of Figure 3. The top curve is of the differences when $v = 2$ and $v = 1$, the bottom for $v = 13$ and 12.

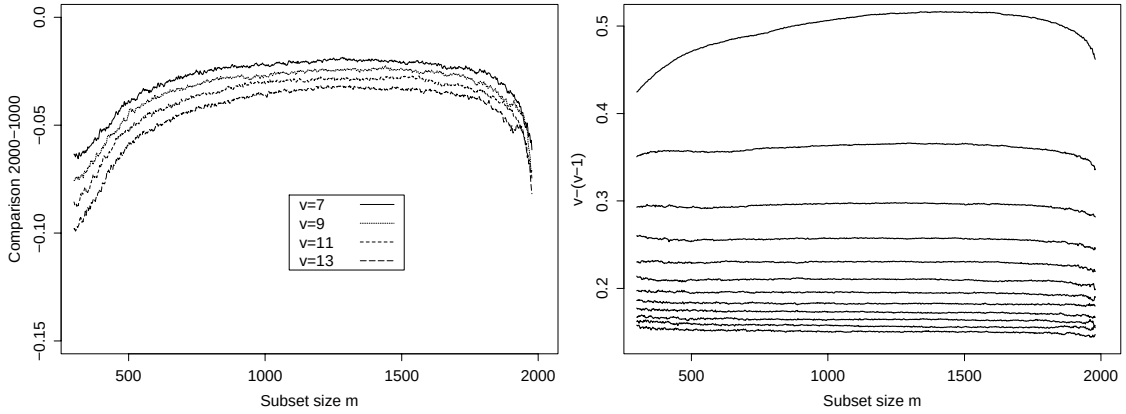

Figure 3: Differences in rescaled minimum Mahalanobis distances of Figure 2. Left-hand panel: curve for n = 2,000 minus that for n = 1,000; v = 7, 9 ,11 and 13. Right-hand panel: differences for v and $v-1$ averaged over the four values of n in Figure 2

These plots show that, for larger values of v, the differences are virtually constant, decreasing as v increases. The greatest departure from this simple structure is the comparison between $v = 1$ and $v = 2$. To elucidate the structure we take average values around the centre of the range. Let these values be

$$\Delta_v = \text{avge}\ (d_v - d_{v-1}), \tag{5}$$

where "avge" is the average, over the centre part of the range, of the differences in the right-hand panel of Figure 3 between the value of the distances for v and $v-1$.

The plot of Δ_v^{-2} against v, not given here, for values in the centre of the figure, is virtually a straight line of slope close to 4. We give an explanation of this value in Appendix 2.

5 Supermarket Data

It seemed from the left-hand panel of Figure 1 that there were several outliers in the supermarket data. The left-hand panel of Figure 4 repeats this plot with the addition of simulation bands. At the end of the curve the observed values lie outside the envelope, indicating some outliers. Confirmatory evidence is provided by the too small values in the centre of the search. In the figure the continuous bands were found by direct simulation and the dotted lines by scaling the curves for n = 2,000 and $v = 13$ to our example with n = 1,827 and $v = 10$. The agreement between the two sets of bands is excellent. The outliers are shown more clearly in the right-hand panel of the figure which is a zoom of the last part of the search.

It is clear that there are several outliers. However, to identify all observations from m = 1,791, where the observed values first lie outside the envelope, as outliers would be to ignore the rapidly increasing shape of the envelope towards the end of the search for all sample sizes. We therefore try the effect of rejecting a few outliers, recalculating the approximate envelope for each new subsample size.

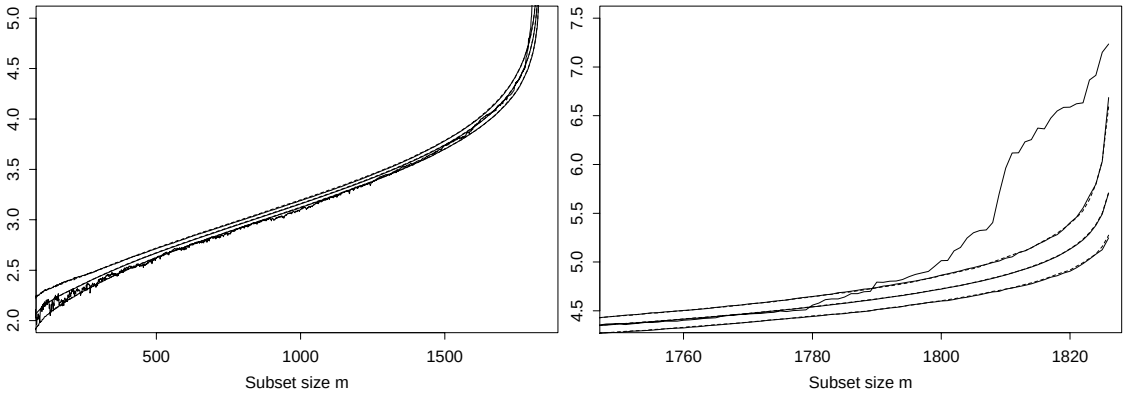

Figure 4: Supermarket data. Left-hand panel, forward plot of minimum scaled Mahalanobis distance with, continuous lines, 1, 50 and 99% points from 10,000 simulations of 1,827 observations and, dotted lines, approximate bounds from rescaling. Right-hand panel, the last part of the search. How many outliers are there?

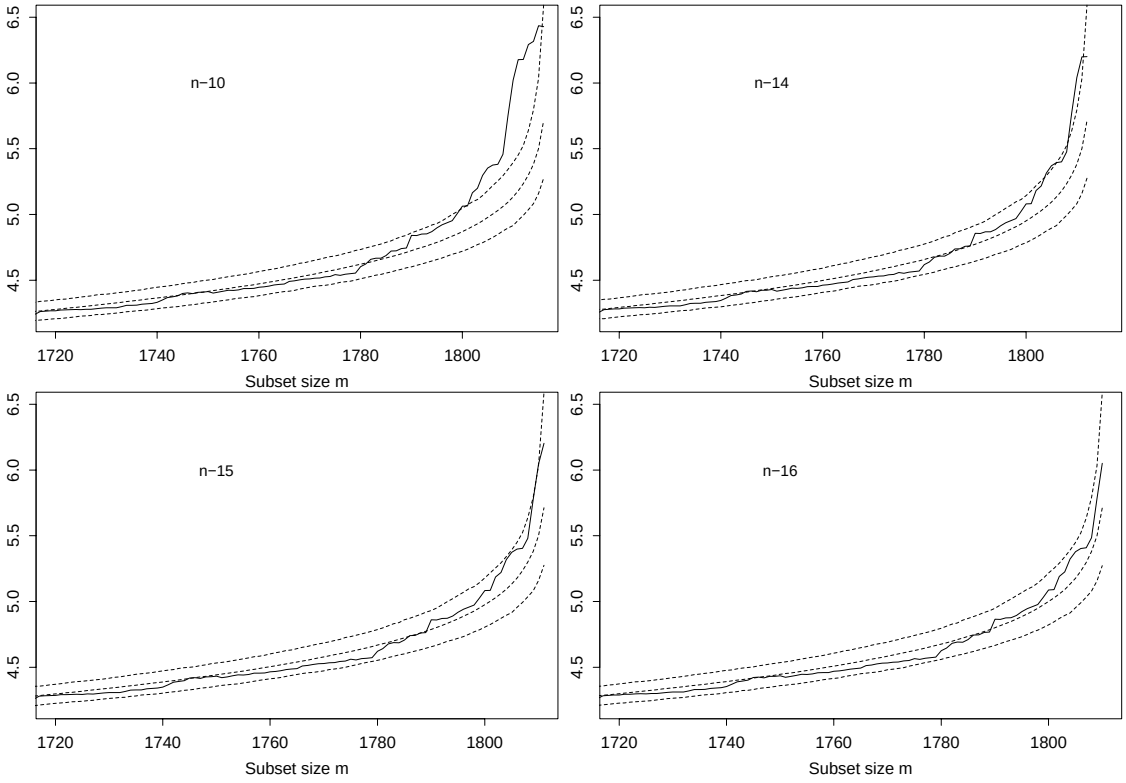

Figure 5: Supermarket data: the last part of the forward plot of minimum scaled Mahalanobis distance as in Figure 4. Successively the last 10, 14, 15 and 16 observations have been deleted. There are 16 outliers

Figure 5 shows the end of the search and the recalculated envelopes when the last 10, 14, 15 and 16 observations are treated as outliers. The lower right-hand panel now shows that when these 16 observations have been deleted, there is no evidence of any further outliers. When we refer our inference back to scatterplots of the data, not given

here for lack of space, we see that working back in the search trims those values that are extreme, leading to a set of observations with elliptically contoured two-dimensional marginal distributions.

The analysis of these data using unscaled distances leads to conclusions that are identical to those found here. Additional plots, including those for this parallel analysis, are available at `www.riani.it/fmvo`.

Appendix 1: Supermarket Data

The data are taken from 1,827 records extracted from loyalty cards for a supermarket chain in Italy. They are part of a larger set of readings from which we have selected 10 variables. The data are available at `www.riani.it/fmvo`. Expenditures are in euros.

y_1: Amount spent
y_2: Amount spent on supermarket's own brands
y_3: Amount spent on products with the lowest price in their category
y_4: Total number of items bought
y_5: Total number bought of supermarket's own brands
y_6: Total number of promotions joined by the customer
y_7: Number of receipts
y_8: Average amount spent per receipt y_1/y_7
y_9: Amount spent in sectors with a high mark up
y_{10}: Percentage of expenditure in sectors with a high mark up out of total expenditure 100 y_9/y_1

The variable y_7 measures the number of visits on which a purchase was made and the loyalty card was brought with the customer.

We removed from our data all customers with zeroes in y_2 and y_3, y_5 and y_6 and y_9 and y_{10}. Scatterplots show that the variables are highly skewed. The methods for robust multivariate transformations in Chapter 4 of [3] indicate transformations close to the logarithmic. This transformation gave variables for which the bulk of the data appeared to be close to the multivariate normal. For numerical stability we jittered the data by adding standard normal random variables multiplied by 10^{-10}.

Appendix 2

To find an approximate expression for Δ_v (5) we use the general structure of the curves, that move up or down without much change in shape as v varies. In the null case the curves are of order statistics from a χ^2 distribution. We consider the properties of χ^2 random variables, the order statistics in the figures being scaled in a similar way as v varies as the variables themselves.

Let $X_v^2 \sim \chi^2_v$. Then $\mathrm{E}\,(X_v^2) = v$. The quantity of interest Δ_v is the expectation of a random variable

$$D_v \doteq \sqrt{X_v^2} - \sqrt{X_{v-1}^2}.$$

If we make the crude approximation that $\mathrm{E}\{g(X)\} = g\{\mathrm{E}(X)\}$,

$$\Delta_v = \mathrm{E}(D_v) = \sqrt{v} - \sqrt{v-1} \quad \text{and so} \quad 1/\Delta_v = \sqrt{v} + \sqrt{v-1}.$$

Then

$$1/\Delta_v^2 = 2v - 1 + 2v\left(1 - \frac{1}{v}\right)^{1/2}.$$

Taylor expansion for large v then shows that

$$1/\Delta_v^2 = 4v - 2 + O(1/v),$$

close to the results at the end of §4.

References

[1] A. C. Atkinson. Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, 89:1329–1339, 1994.

[2] A. C. Atkinson and M. Riani. Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis*, 2007. doi:10.1016/j.csda.2006.12.034.

[3] A. C. Atkinson, M. Riani, and A. Cerioli. *Exploring Multivariate Data with the Forward Search*. Springer–Verlag, New York, 2004.

[4] L. Davies and U. Gather. The identification of multiple outliers (with discussion). *Journal of the American Statistical Association*, 88:782–801, 1993.

[5] W. K. Fung. Letter to the editor: Critical values for testing multivariate statistical outliers. *Applied Statistics*, 45:496–497, 1996.

[6] A. S. Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B*, 54:761–771, 1992.

[7] J. Hardin and D. M. Rocke. The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14:910–927, 2005.

[8] N. Neykov, P. Filzmoser, R. Dimova, and P. Neytchev. Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics and Data Analysis*, 2007. doi:10.1016/j.csda.2006.12.024.

[9] P. J. Rousseeuw and B. C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–9, 1990.

[10] S. S. Wilks. Multivariate statistical outliers. *Sankhya A*, 25:407–426, 1963.

[11] D.L. Woodruff and D. M. Rocke. Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association*, 89:888–896, 1994.